University of the Witwatersrand, Johannesburg
School of Family Medicine

FAMH4004A/COMS5027A

Health Analytics

Lab Assessment 5: k-means clustering

Date, 2024

Consultation Times

Description

In the lectures we have covered k-means clustering as our first unsupervised learning model. We have covered the simple operations of the algorithm.

For this assessment, you will use the *"airpoll"* dataset and run k-means clustering. In addition, you must also write a report of roughly 3-5 pages describing your implementation, justifying your decisions and presenting your results. You are allowed to use Numpy, Pandas, Matplotlib, Seaborn and sklearn. Other libraries can be used. Please see the rubric which is presented below and can also be found in the course outline for an indication of how marks will be allocated.

Note that it is not sufficient to just implement all possible concepts we have covered for an arbitrary dataset. We are concerned with your thought process and decision-making. In the real world, you will usually be presented with a dataset and have to figure out which model to use. In this lab, however, you must use a linear regression.

In your assignment, you should showcase the following

- Description of your variables
- Show the correlation among the variables
- Show the distribution of each pollutant
- Show the ideal number of clusters for the k-means run
- Show the correct process to execute k-means clustering
- Export results to an Excel spreadsheet

uLWAZI Forum

We will be using a forum on uLWAZI to engage with each other and the lecturer/tutors outside of consultation. This is an important aspect of the course, as in the practice a lot of how we learn to implement ML and DS algorithms is by asking questions and reading on forums. Thus, a portion of the course mark will be attributed to how you engage on the forum and the quality of your engagements. Please read the following carefully as we will be adhering to the StackOverflow (a real coding forum) standards on the forum:https://stackoverflow.com/help/how-to-ask. Since we are all working on different datasets, discussion is encouraged. Abusive or hurtful behaviour on the forum will not be tolerated. Nor will out-rightly giving away answers or code. As is the StackOverflow standard you may give clear demonstrations, particularly using the minimal working examples provided with the question, but answers cannot contain full solved problems and asking other to do your work is also not allowed.

Submission

Please upload your code (in a zip file) and the written report to the uLWAZI assignment portal. Your code should be reproducible from just what is uploaded, so please include the data and do not use any packages other than the three standard ones. Please also refrain from submitting exceptionally large datasets (100Mb or more). It is advised that you keep things rather simple and focus more on a thorough application of the concepts taught in class, rather than working on difficult data or complex settings which require a lot of domain knowledge.

Submission date and time:17th November,2024 at 23:59.

Academic Integrity

We take academic integrity extremely seriously. Communication during quizzes, accessing someone else’s Moodle account, plagiarism and dishonesty etc. will be dealt with decisively according to the school and University procedures. You will receive 0 for the assessment, may receive FCM for the module, and may be reported to the legal office where relevant. During online learning, it is particularly important to maintain the standards expected at Wits. The lecturer may, at his discretion call on you to replace any assessment with an equivalent oral presentation